Naval Research Laboratory

Washington, DC 20375-5000

NRL Report 9160

Vocabulary Synthesis Based on Line Spectrum Pairs

STEPHANIE S. EVERETT

Human-Computer Interface Laboratory
Information Technology Division

January 12, 1989

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Report 9160	5 MONITORING ORGANIZATION REPORT NUMBERS
6a NAME OF PERFORMING ORGANIZATION Naval Research Laboratory	6b OFFICE SYMBOL (If applicable)
6c ADDRESS (City, State, and ZIP Code) Washington, DC 20375-5000	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING/SPONSORING ORGANIZATION	8b OFFICE SYMBOL (If applicable)
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61153N; PROJECT NO: RR021-05-42; TASK NO: ; WORK UNIT ACCESSION NO: DN2573

11 TITLE (Include Security Classification)
Vocabulary Synthesis Based on Line Spectrum Pairs

12 PERSONAL AUTHOR(S)
Everett, Stephanie S.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Interim	FROM 12/86 TO 12/87	1989 January 12	16

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Speech synthesis; Human-computer interface; Text-to-speech; LSP

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This report describes the initial investigation of a new synthetic speech system based on a line spectrum pair (LSP) representation of the speech spectral envelope. The system contains a library of stored LSP speech segments extracted from natural speech. These segments are modified as necessary by a small set of context-sensitive rules and then concatenated to generate high-quality, natural-sounding speech. Tests of a preliminary system produced Modified Rhyme Test and Diagnostic Rhyme Test scores of 87.3 and 79.7, respectively. This LSP vocabulary synthesizer is currently limited to single syllable utterances, but future research will expand its capabilities to allow implementation of a full text-to-speech system.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL Stephanie S. Everett	22b TELEPHONE (Include Area Code) (202) 767-2116
	22c OFFICE SYMBOL Code 5531

DD Form 1473, JUN 86 *Previous editions are obsolete* SECURITY CLASSIFICATION OF THIS PAGE

CONTENTS

VOCABULARY SYNTHESIS BASED ON LINE SPECTRUM PAIRS

BACKGROUND

Currently there are two primary types of unlimited vocabulary speech synthesis systems available. They are: formant-based synthesis by rule systems and linear-predictive-coding- (LPC-) based segment concatenation systems. The first synthesis-by-rule system appeared in the early 1960s [1]. In the more than 20 years since that pioneering effort, many other systems have been developed [2-5, and others, see Ref. 6 for a review] and the parameter generation rules have been greatly refined [7,8]. However, the basic approach has remained the same: a large set of rules is used to generate the parameters (formant frequencies, bandwidths, and amplitudes) needed to drive a bank of bandpass filters. A great deal has been learned about the relationships between formant trajectories and speech sounds, and the intelligibility of the newer systems is often quite good. However, they all suffer from the flat, nasal sound characteristic of the formant synthesizer itself. This mechanical *accent* is due, in part, to the fact that the synthesizer requires a fixed number of formants (usually three) and the trajectories must be continuous, whereas in natural speech the number of formants can vary from one to five and the trajectories may not be continuous.

The other type of synthesizer available is based on the concatenation of stored segments excised from natural speech [9-12]. The individual segments are stored as sets of parameters derived by LPC analysis of the original speech signal. Because the LPC synthesizer is sensitive to discontinuities or rapid changes in parameter trajectories, the parameters must be smoothed or interpolated across concatenation boundaries. However, this smoothing can degrade the quality and intelligibility of the synthesized speech. Furthermore, an error in any one LPC parameter affects the entire speech spectrum. Another drawback to this type of system is that LPC is not adequate for representing certain sounds such as voiced fricatives, which require a mixed excitation source (both voiced and unvoiced simultaneously), or voiced stops, which require very rapid transitions.

This report describes the initial development of a new vocabulary synthesizer based on the line spectrum pair (LSP) approach to speech processing [13]. The use of LSPs as speech parameters offers several advantages over the existing approaches to unlimited vocabulary synthesis. Though they are directly related to LPC parameters, LSPs are frequency-domain parameters like formants, so the vast body of knowledge about speech spectral properties can be used in the development of this system. Unlike LPC parameters, LSPs are very tolerant of error—an error in one LSP affects the spectrum only in that frequency region. LSPs are also tolerant of rapid changes in trajectory, so no smoothing or interpolation is required at concatenation boundaries. Unlike formant synthesizers, the LSP synthesizer is capable of producing highly natural speech quality, and the excitation signal is well defined (it is the prediction residual). The number of LSP frequencies is constant and naturally ordered, and the trajectories are continuous, even across unvoiced sounds.

The LSP synthesizer contains a library of approximately 150 basic speech segments excised from natural speech and stored as sets of LSP parameters. For each utterance, the specified segments are retrieved from the library, scanned by a small set of context-sensitive rules, and modified if necessary. Pitch and amplitude curves are computed, and the concatenated segments are then output through the LSP synthesizer.

Manuscript approved August 1, 1988.

LSP Analysis

LSP parameters are derived from LPC prediction coefficients through the decomposition of the impulse response of the LPC analysis filter $A(z)$ into even and odd functions. The transfer function of $A(z)$ may be expressed as

$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_n z^{-n}, \tag{1}$$

where a_n is the n^{th} prediction coefficient. By taking a sum and difference between $A(z)$ and its conjugate function (the transfer function of the filter whose impulse response is the mirror image of $A(z)$), $A(z)$ can be decomposed to a sum of two filters $P(z)$ and $Q(z)$, each having roots along the unit circle of the complex z plane:

$$P(z) = A(z) - z^{-(n-1)} A(z^{-1}) \tag{2}$$

and

$$Q(z) = A(z) + z^{-(n-1)} A(z^{-1}). \tag{3}$$

Figure 1 illustrates this decomposition. $A(z)$ can then be reconstructed by using the sum of these two filters as

$$A(z) = \frac{1}{2}[P(z) + Q(z)]. \tag{4}$$

The roots of $P(z)$ and $Q(z)$ can be found by searching for null frequencies in their amplitude spectra. The estimated line spectrum frequency is refined through a simple parabolic approximation based on the three consecutive spectral points nearest the null frequency.

Figure 2 shows formant and LSP trajectories from an actual speech sample. Closely spaced lines correspond to speech resonant frequencies; widely spaced lines correspond to valleys in the speech spectral envelope.

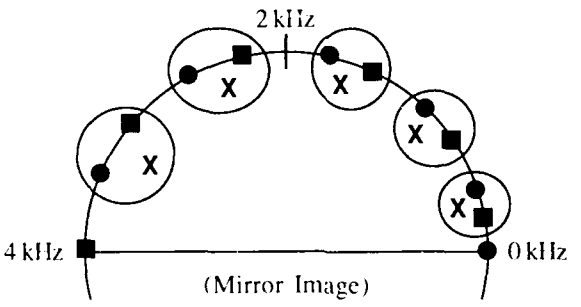


Fig. 1 — Decomposition of the roots of a tenth-order LPC analysis filter with a speech sampling rate of 8 kHz. Each root of $A(z)$, indicated by **X**, is decomposed into two roots indicated by ● and ■. Related roots are circled.

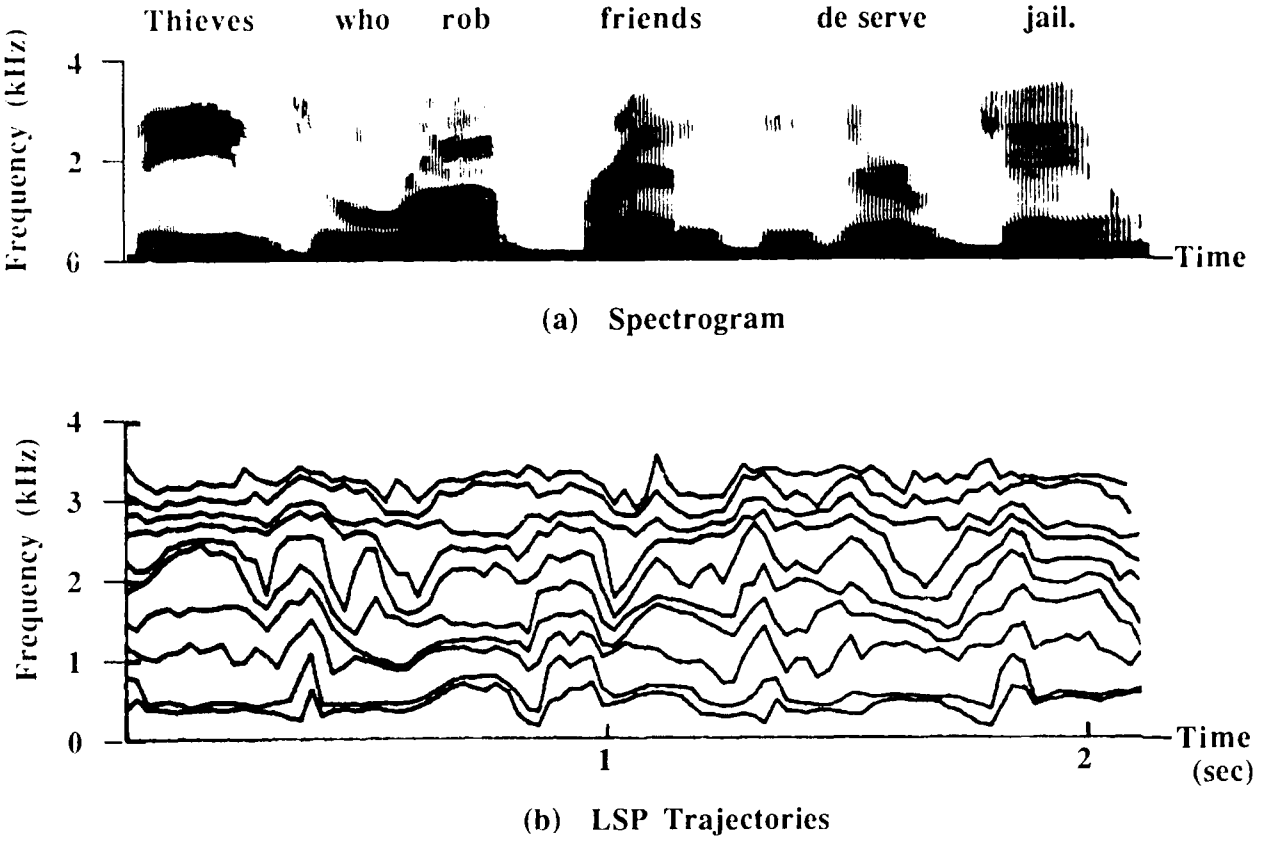


Fig. 2 — Typical LSP trajectories, with a spectrogram of the original speech showing the formant trajectories for comparison. Because LSP parameters follow the speech resonant frequencies, the trajectories are very similar. Note that the number of LSP frequencies is constant and naturally ordered, and that the trajectories are continuous, even across unvoiced sounds.

LSP Synthesis

The LSP synthesis is the inverse of the LSP analysis and is very similar to LPC synthesis. Speech may be generated directly from LSPs or by converting the LSPs back to the prediction coefficients used in LPC synthesis; this system uses the LSPs directly. Four enhancements were made to the standard LSP/LPC synthesis algorithm for the vocabulary synthesizer in this report [14,15]. These were the introduction of random components into the voiced excitation source, the modification of the unvoiced excitation signal at the onsets of unvoiced stops, the use of a soft voicing decision, and the expansion of the output speech bandwidth to approximately 6.5 kHz.

In the standard synthesizer (either LSP or LPC), the waveform of the voiced excitation signal repeats exactly from one pitch cycle to the next. In contrast, the prediction residual rarely repeats exactly from one pitch cycle to the next. This is due to irregularities in vocal cord movement and turbulent air flow from the lungs during the glottis-open period of each pitch cycle. The extreme regularity of the standard excitation signal causes the synthesized speech to sound machinelike and tense. To reduce this effect, pitch epoch variations and period-to-period waveform variations have been realized by introducing phase jitter into the voiced excitation signal waveform. This gives the synthesized speech a pleasant breathy quality, and the buzzy, twangy sound usually associated with LPC is greatly reduced.

The excitation signal traditionally used for unvoiced sounds is uniformly distributed random noise. This is satisfactory for the reproduction of fricatives (*/s/*, */f/*, */th/*, */sh/*), but not for plosives

(/p/, /t/, /k/, /ch/). Because the sudden burst at the onset of a plosive cannot be predicted well by the past speech samples, plosives often sound more like fricatives. To alleviate this problem, randomly spaced pulses have been added to the conventional unvoiced excitation signal in this synthesizer. The amplitude of the pulses is proportional to the abruptness, or change in speech energy (rms), at the beginning of the sound. At the relatively gentle onsets of fricatives, the pulses are insignificant. At the sudden onsets of plosives, however, the randomly spaced pulses comprise a major portion of the excitation signal. This results in the production of much more robust and intelligible plosives.

Rather than the traditional mutually exclusive voiced/unvoiced decision, the soft voicing decision used in this system determines the degree of voiced and unvoiced excitation in each sound. It is based on the fact that whenever there is periodic excitation (i.e., voicing), there is a spectral resonance in the lower frequency regions (a first formant), and the first LSP is therefore low and narrow. When there is no periodic excitation (i.e. the sound is unvoiced), there is no spectral resonance in the lower frequency region, and the first LSP is higher and wider. Therefore, by using the first LSP, it is possible to derive a soft voicing indicator β such that $0 \leq \beta \leq 1$. This indicator is used as a weight in the computation of the mixed excitation source E_m, with

$$E_m = \beta E_v + (1 - \beta) E_{uv}, \tag{5}$$

where β is the voicing indicator, E_v is the voiced excitation signal, and E_{uv} is the unvoiced excitation signal. Use of the mixed excitation source improves the reproduction of complex sounds such as voiced fricatives and stops.

The voice quality of the standard synthesizer is often described as muffled. This is because the output speech bandwidth is only 4 kHz for a system that samples the input speech at 8000 samples per second, whereas the bandwidth of natural speech exceeds 8 kHz. To help improve the quality of the synthesized speech, the output bandwidth of this system is expanded by using a two-step postsynthesis process. First, a zero is inserted between each pair of adjacent samples in the output digital waveform, doubling the data rate (from 8 to 16 kHz). This effectively reflects the spectrum in the region from 0 to 4 kHz upwards into the 4 to 8 kHz region. The signal is then low-pass filtered at approximately 6.5 kHz to remove unwanted components at the higher frequencies. This expanded output bandwidth makes the synthesized speech sound much brighter and less muffled.

TECHNICAL APPROACH

The vocabulary synthesizer described here contains a set of approximately 150 segments excised from natural speech and stored as LSP parameters. For each utterance, the specified segments are retrieved from the library, scanned by five context-sensitive rules, and modified if necessary. Pitch and amplitude curves are computed, and the concatenated segments are then output through the LSP synthesizer. Figure 3 summarizes this process.

LSP Segment Library

A list of over 200 words was read by one male speaker and then digitized at an 8 kHz sampling rate using a 4 kHz low-pass anti-aliasing filter. The speech was passed through the LSP analysis (tenth order, 130 samples per frame), and the parameters for each word were stored in a separate file. The LSP trajectories were then plotted, and the desired segments were carefully excised and placed in the library. Individual parameters and trajectories were sometimes adjusted manually to improve the segments by minimizing unwanted coarticulation, removing irregularities, etc.

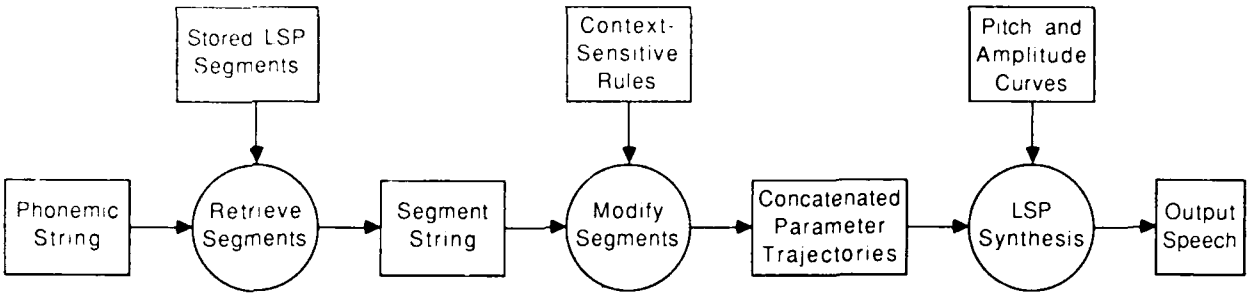


Fig. 3 — Summary of the Naval Research Laboratory (NRL) vocabulary synthesis system

This synthesizer differs from most existing ones in that it does not use the traditional fixed-size units found in most concatenation systems. The excised segments in this system range in size from subphonemes, such as the *w*-like off-glide in *how*, to full syllables, such as *eel* and *or*. The use of segments of varying lengths reduces the need for complex combinatory rules but still produces high-quality speech. For example, post-vocalic *L* and *R* have very strong coarticulation effects on the preceding vowel, so the vowel-liquid sequence is simply stored as a unit, rather than trying to adjust the parameter trajectories by rule. Likewise, two segments are stored for most initial stops: one for use before front vowels, and one for use before nonfront vowels. This ensures efficient synthesis of natural-sounding speech, since the coarticulation present in the original speech is preserved within each segment.

Each segment in the library contains a header that includes the following information:

- segment length—the number of frames of LSP parameters stored in the file;
- subsegment length—the length in frames of a shorter version of the segment, used for vowels before unvoiced consonants and for consonants in clusters;
- segment type—voiced or unvoiced;
- phonemic class—vowel, consonant (fricative or nonfricative), or liquid/glide;
- position of articulation—front, medial, or back;
- height of articulation—high, middle, or low.

The segment lengths, voicing, and phonemic class information are used to determine the applicability of the context-sensitive concatenation rules. Articulation information is not required at the current time but has been included for future use with text analysis.

Concatenation Rules

This synthesis system uses three context-sensitive concatenation rules to adjust segment lengths in certain environments. Other synthesis-by-concatenation systems handle contextual segment variation by repeating parameters [16], by repeating segments [17], by time-warping [18], or with a pronouncing dictionary [9,11].

Two additional rules smooth the amplitude and voicing curves at the onset of initial vowels and the offset of final vowels. Note that this smoothing is included to improve the naturalness of the output speech and is not required by the synthesizer. All of the rules are described in detail below.

Rule 1—Use the short version of a vowel segment preceding an unvoiced consonant. This rule handles the adjustment of vowel length that is one of the primary indicators of voicing in word-final consonants. The longer version for use before voiced consonants is stored in the library. A shorter subsection of this segment is used before unvoiced consonants.

Rule 2—Lengthen a vowel preceding a fricative consonant. This rule reflects the fact that in isolated words, vowels before fricatives are approximately 30% longer than vowels before plosives [19,20]. (The segment library for this synthesizer contains separate vowel segments for use before nasals, liquids, and glides.) Because vowel length is also dependent on the voicing of the final consonant (see Fig. 4), Rules 1 and 2 must be applied in order.

Rule 3—Use the short version of a consonant when it is followed by another consonant segment. This adjusts the length of nonfinal consonants in postvocalic clusters (e.g., *act*, *send*, *desks*). In cases where the following consonant is a fricative, the subsegment length is increased by one frame to include the release of stop consonants at the transition to the fricative (e.g., *rapt* vs *raps*).

Rule 4—Smooth the onset voicing and amplitude curves of initial vowels. If the first frame of a vowel is produced at full strength and full voicing, the abrupt onset can give the impression of an initial stop and sometimes generates a pop in the synthesizer. This rule produces a gentler, more natural vowel onset by smoothing the amplitude curve in the first two frames and by increasing the amount of unvoiced excitation in the first frame.

Rule 5—Smooth the offset voicing and amplitude curves of final vowels. This rule lengthens word-final vowels by two frames and produces a natural-sounding vowel offset. The first added frame has the same filter parameters as the preceding frame, but is lower in amplitude and has a higher proportion of unvoiced excitation in the excitation signal. In the final frame (the second added frame), the LSP trajectories move toward a neutral position (i.e., a flat spectrum), the amplitude decreases further, and the excitation signal is mostly unvoiced.

Each segment in the utterance is processed in sequence in a single pass from left to right. The rules are applied in order, by using the information in the headers of the current and following segments to determine the applicability of each rule.

Pitch and Amplitude Curves

Because the capabilities of this vocabulary synthesizer are currently limited to utterances of isolated single-syllable words, the pitch trajectory is simply a linear function of time. For each word, the pitch drops linearly from approximately 125 Hz at the beginning of the word to 80 Hz at the end.

The segments in the LSP library were excised from naturally spoken single-syllable words, so the shapes of the original amplitude curves were maintained. Actual amplitude values were retained for (most) consonants; vowels are normalized to a specified maximum amplitude. These normalized values are used as weights and are superimposed on an overall amplitude curve that is a linear function of time. In this way, the relative loudness of the vowels is maintained (e.g., *ow* is inherently louder than *eh*).

INTELLIGIBILITY TESTS

Two tests were conducted to measure the intelligibility of the synthetic speech. The first was a Diagnostic Rhyme Test (DRT), which tests the intelligibility of word-initial consonants by using a set of 224 single-syllable words. The words are paired, with the members of each pair differing only in

one consonantal attribute (e.g., *goat-coat* [voicing]; *nip-dip* [nasality]). A single DRT list was synthesized and scored by eight trained listeners. The overall score was 79.7% correct (corrected for guessing). Table 1 breaks down this score.

The second test conducted was a Modified Rhyme Test (MRT), which tests the intelligibility of both initial and final consonants by using 50 sets of 6 words (e.g., *bill-hill-fill-will-kill-till*; *mass-map-math-man-mad-mat*). Twenty subjects each listened to 6 test lists preceded by a training list of 50 items chosen at random from the 300 test items. Subjects' experience with processed speech varied considerably, but this did not affect their performance ($t = 0.012$, $p < .001$). Overall accuracy on the MRT was 87.3% (85.2% corrected for guessing). Accuracy on initial contrasts was 82.3%; accuracy on final contrasts was 92.0%. Final consonants are generally easier to distinguish because significant information is conveyed by the length and nasality of the preceding vowel, and by coarticulation at the vowel-consonant transition.

Table 1 — DRT Results for the NRL Vocabulary Synthesis System (Numbers Indicate Percent Correct)

Attribute	Present	Absent	Mean
Voicing	89.1	98.4	93.8
Nasality	70.3	93.8	82.0
Sustention	39.1	53.1	46.1
Sibilation	98.4	96.9	97.7
Graveness	70.3	70.3	70.3
Compactness	92.2	84.4	88.3
Total		79.7	

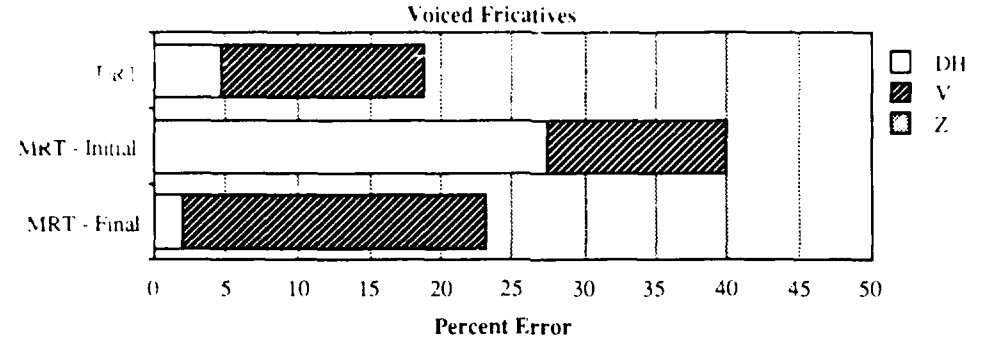
Table 2 and Fig. 4 summarize the DRT and MRT results by sound class. Because the two tests are quite similar, their scores are highly correlated, and the patterns of errors should be reasonably consistent between tests. However, because the DRT offers only two alternatives for each item, whereas the MRT offers six, the error rates on the MRT are apt to be somewhat higher.

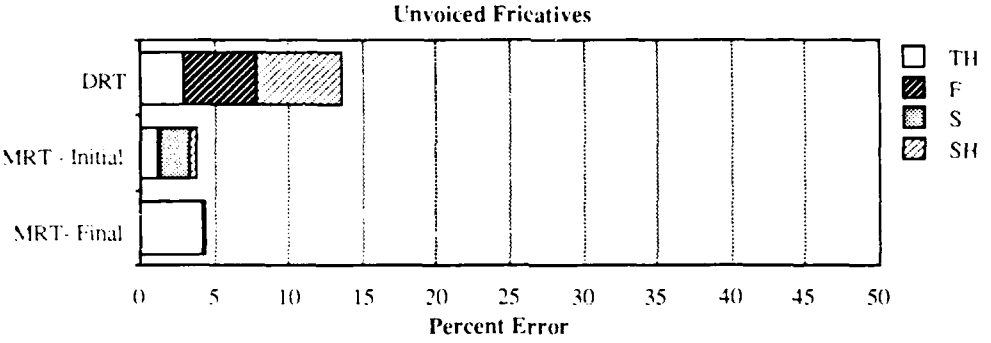
Table 2 — Analysis of Intelligibility Test Results

Sound Class[a]	DRT		MRT Initial Contrasts		MRT Final Contrasts	
	# Contrasts[b]	% Error	# Contrasts[b]	% Error	# Contrasts[b]	% Error
Stops & Affricates						
Unvoiced /p, t, k, č/	42	4.5	34	8.1	46	4.8
Voiced /b, d, g, dž/	64	10.0	32	42.8	25	5.6
Fricatives						
Unvoiced /f, θ, s, š/	36	13.5	30	3.7	21	4.3
Voiced /v, ð, z/	16	18.8	2	40.0	10	23.0
Nasals /m, n, ŋ/	19	13.8	12	17.1	32	16.3
Liquids & Glides /l, w, r, y, h/	15	1.7	38	12.0	13	0.4
No consonant	0		2	40.0	3	0.0
TOTAL	192	10.2	150	17.7	150	8.0

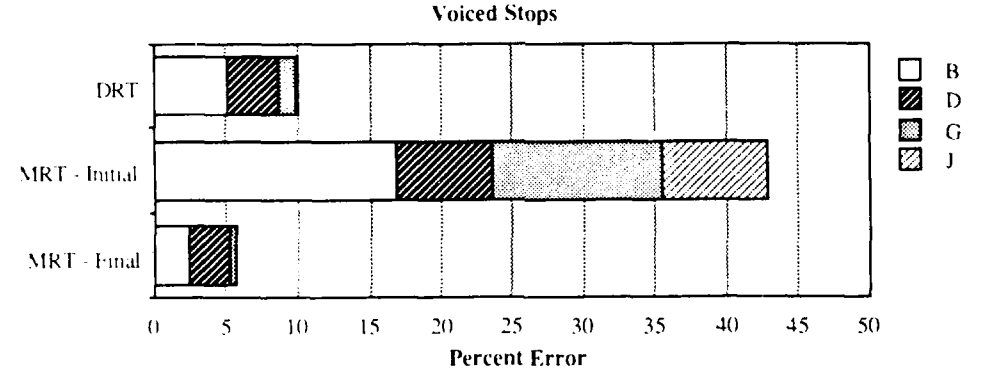
[a] The six phonetic categories used on the DRT are not applicable to the MRT because the MRT contains multidimensional contrasts.

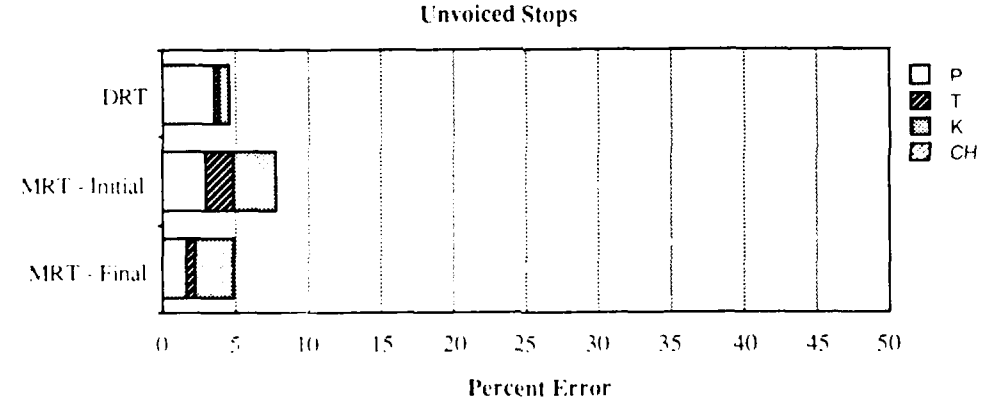
[b] The number of times a consonant of that class appears in the test list.





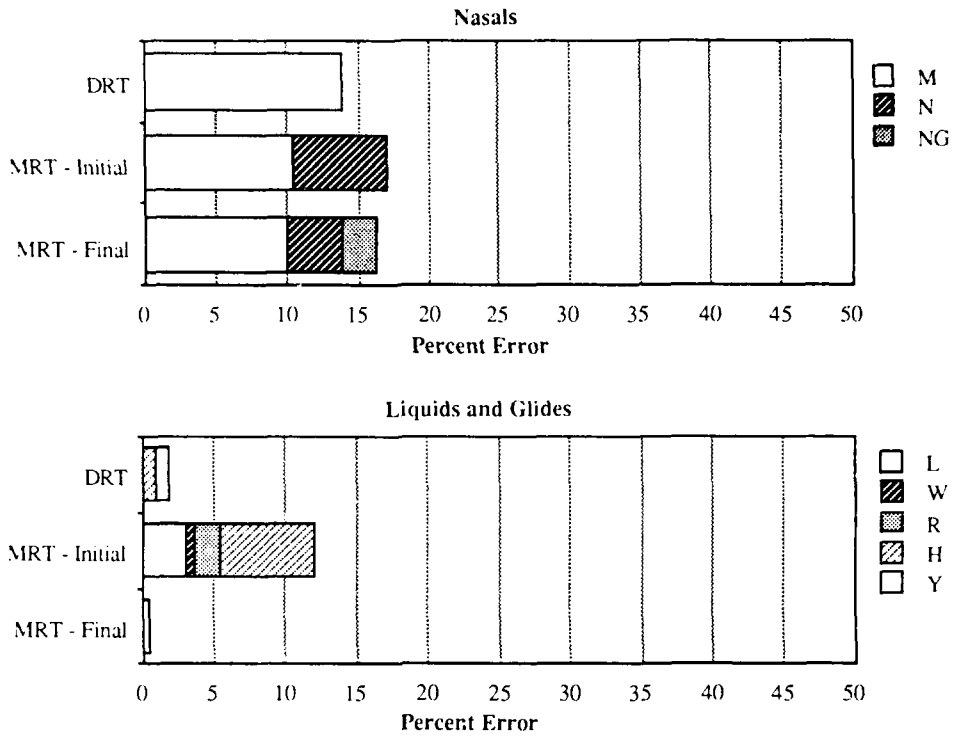
(a) Voiced and unvoiced fricative segments





(b) Voiced and unvoiced stop segments

Fig. 4 — Distribution of intelligibility test scores



(c) Nasal, liquid, and glide segments

Fig. 4 (Cont'd) — Distribution of intelligibility test scores

Overall, the highest error rates on both tests occurred with the voiced fricatives, specifically *V* and *TH* (as in *then*); the error rate for *Z* on both tests was 0.0%. On the DRT, the error rate for *V* and *TH* was 18.8%. Of these errors, 92% were confusions with the voiced stops *B* and *D*, and 8% with the unvoiced fricatives *F* and *TH*. On the MRT, where the error rate for voiced fricatives was 40%, 13% of the errors on *V* and *TH* were confusions with *B* and *D*, and 69% with *M* and *N* (the DRT does not test *V* vs *M* or *TH* vs *N*). These errors may be caused by improper amplitude curves at the vowel-consonant transitions, or by too much periodic excitation in the mixed excitation signal for these segments. It is also possible that new or additional segments are needed for these sounds.

On the DRT, the errors for unvoiced fricatives fell into two areas. The first was the segment *SH*, with an overall error rate of 37%, which accounted for 42% of the errors in this class. All of these errors occurred with the *SH* − *CH* contrast. This is a problem of sustention, and indicates that the onset amplitude curve of *SH* is too abrupt. The segment may also be too short. The *F* − *TH* distinction accounts for another 44% of the unvoiced fricative errors on the DRT. The error rate on this contrast was 59%. These sounds are very similar, and the *F* − *TH* distinction is always difficult, even with natural speech. These errors may therefore result more from limitations inherent in the speech processing (particularly the restricted bandwidth) than from problems with the segments themselves. Unfortunately, neither the *SH* − *CH* nor the *F* − *TH* contrast is tested on the MRT. On the MRT, the error rate for *SH* was very low—it was misidentified only twice, both times as *F*. However, the error rate for *TH* was relatively high. In initial position, *TH* was heard as *P* 35% of the time, accounting for 32% of the errors in this class. (Another 54% of the errors in this class were cases where *S* was heard as *F*. However, because of the frequent occurrence of this contrast on the MRT, the actual error rate for *S* was only 4%.) In final position, the error rate for *TH* was 17%, accounting for almost 95% of the errors on the unvoiced fricatives. Of these, 72% were confusions with *S*, and 21% with *T*.

On the DRT, all the errors in the nasal class were misidentifications of *M*, where the error rate was 13.8%. Of these errors, 90% were cases where *M* was heard as *B* (43% before front vowels; 57% before back vowels); the remaining 10% were cases where *M* was heard as *N*. On the MRT, the error rates for *M* were 18% and 35.6% in initial and final positions, respectively, most of them confusions with *N*. This accounted for 61% of the errors on initial nasals and 62% of the errors on final nasals. From this, it is evident that the *M* segments are in need of modification or replacement to improve the intelligibility of nasal consonants.

On the MRT, a 40% error rate was obtained on words with no initial consonant (*eel* and *oil*). Most of these errors were confusions between *oil* and *foil* (45%) or *boil* (20%). *Eel* was only occasionally confused with *peel* (10%) and *heel* (5%). These results indicate a problem with the *OIL* segment, and possibly a more general problem with the voicing and amplitude curves at the onsets of initial vowels. The DRT does not include items with initial vowels.

Table 3 compares the intelligibility test scores for this system with published scores for several other systems. Note that these other systems are fully developed commercially available text-to-speech systems, whereas the NRL system is still in the earliest stages of development. It is expected that intelligibility of this system will improve significantly as research progresses. (As a further comparison, an analysis-synthesis communication system using a similar LSP synthesis algorithm scored 93 on the DRT [21].)

Table 3 — Comparision of DRT [22] and MRT [23] Scores for Several Commercial Text-to-Speech Systems and the New LSP-Based System Described in this Report

Voice	DRT	MRT
Natural speech	95.6	99.4
DECtalk Paul	87.5	96.7
DECtalk Betty	92.4	94.4
Prose 2000	81.2	94.3
Infovox	83.6	87.4
Type and Talk	65.9 (Namal)	66.2 (Votrax)
NRL LSP system (Prototype)	79.7	87.3

FUTURE RESEARCH

All of the problem segments discussed above will be carefully evaluated and modified or replaced as necessary to improve intelligibility. Additional segments may be required, such as a *V* before front vowels and a *V* before back vowels, rather than one all-purpose *V* segment. Changes or additions to the concatenation or shaping rules, particularly amplitude shaping, may also improve intelligibility.

The capabilities of this preliminary system will be expanded to include the synthesis of multisyllable words, phrases, and sentences. The synthesizer will then be combined with a text analysis system for the development of a full text-to-speech capability. The system may also be expanded to include a variety of voices by collecting libraries of new segments from additional speakers.

REFERENCES

1. J. Kelly and L. Gerstman, "An Artificial Talker Driven from Phonetic Input," *J. Acoust. Soc. Am. Suppl. 1*, **33** (S35) (1961).

2. R.T. Gagnon, "Votrax Real Time Hardware for Phoneme Synthesis of Speech," Proc. ICASSP '78, Tulsa, OK, pp. 175-178.

3. G.F. Groner, J. Bernstein, E. Ingber, J. Pearlman, and T. Toal, "A Real-Time Text-to-Speech Converter," *Speech Tech.* **1**, 73-76 (1982).

4. D.H. Klatt, "The Klattalk Text-to-Speech System," Proc. ICASSP '82, Paris, France, pp. 1589-1592.

5. J. Allen, S. Hunnicutt, and D.H. Klatt, *From Text to Speech: The MITalk System*. (Cambridge Univ. Press, Cambridge, UK, 1987).

6. D.H. Klatt, "Review of Text-To-Speech Conversion for English," *J. Acoust. Soc. Am.* **82**, 737-788 (1987).

7. R. Carlson, B. Granstrom, and S. Hunnicutt, "A Multi-Language Text-to-Speech Module," Proc. ICASSP '82, Paris, France, pp. 1604-1607.

8. S. Hertz, J. Kadin, and K. Karplus, "The Delta Rule Development System for Speech Synthesis from Text," Proc. ICASSP '85, Tampa, FL, pp. 1589-1601.

9. J. Olive and M. Liberman, "A Set of Concatenative Units for Speech Synthesis," in *Speech Comm. Papers*, J.J. Wolf and D.H. Klatt, eds. (Acoust. Soc. Am., New York, 1979) pp. 515-518.

10. J.B. Lovins, M.J. Macchi, and O. Fujimura, "A Demisyllable Inventory for Speech Synthesis," in *Speech Comm. Papers*, J.J. Wolf and D.H. Klatt, eds. (Acoust. Soc. Am., New York, 1979), pp. 519-522.

11. J.P. Olive, "Rule Synthesis of Speech from Dyadic Units," Proc. ICASSP '77, Hartford, CT, pp. 568-570.

12. C.P. Browman, "Rules for Demisyllable Synthesis using Lingua, a Language Interpreter," Proc. ICASSP '80, Denver, CO, pp. 561-564.

13. G.S. Kang and L.J. Fransen, "Experimentation with Synthesized Speech Generated from Line-Spectrum Pairs," *IEEE Trans. ASSP* **ASSP-35**, 568-571 (1987).

14. G.S. Kang and S.S. Everett, "Improvement of the Excitation Source in the Narrowband Linear Predictive Vocoder," *IEEE Trans. ASSP* **ASSP-30**, 377-386 (1985).

15. G.S. Kang and S.S. Everett, "Improvement of the Narrowband Linear Predictive Coder, Part 2: Synthesis Improvements," NRL Report 8799, June 1984.

16. J. Olive, "A Scheme for Concatenating Units for Speech Synthesis," Proc. ICASSP '80, Denver, CO, pp. 568-571.

17. J. May, "Allophone Speech Synthesis Technique," General Instruments Microelectronics Appl. Rept, 1982.

18. R. Schwartz, J. Klovstad, J. Makhoul, D. Klatt, and V. Zue, "Diphone Synthesis for Phonetic Vocoding," Proc. ICASSP '79, Washington, DC, pp. 891-894.

19. G. Peterson and I. Lehiste, "Duration of Syllable Nuclei in English," *J. Acoust. Soc. Am.* **32**, 693-703 (1960).

20. N. Umeda, "Vowel Duration in American English," *J. Acoust. Soc. Am.* **58**, 434-445 (1975).

21. G.S. Kang, unpublished data.

22. R.L. Pratt, "Quantifying the Performance of Text-to-Speech Synthesizers," *Speech Tech.* 54-64, (1987).

23. D.B. Pisoni, H.C. Nusbaum, and B.G.Greene, "Perception of Synthetic Speech Generated by Rule," *Proc. IEEE* **73**, 1665-1676 (1985).